DOCUMENT RESUME

ED 244 594 IR 011 115

AUTHOR Buckland, Lawrence F.
TITLE Data Input for Publishers. State-of-the-Art Report.
INSTITUTION Inforonics, Littleton, Mass.
PUB DATE 8 Mar 84
NOTE 10p.
PUB TYPE Reports - Descriptive (141) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Character Recognition; *Cost Estimates; Guidelines; *Microcomputers; Online Systems; Optical Scanners; *Publishing Industry; Standards; *Word Processing; *Writing for Publication
IDENTIFIERS Data Editing; *Data Entry

ABSTRACT

Intended to serve as a guide for selecting the most suitable automated method for preparing and sending papers for publication, this report briefly describes data entry and editing procedures available from one publisher (Inforonics) and provides an estimate of processing costs for each procedure. The procedures, arranged in order of increasing equipment cost, include optical character recognition (OCR), the use of microcomputers and home computers, utilization of a personal computer supplied by Inforonics and Inforonics' data input software known as TexTerm, data entry using word processors, use of a dedicated microcomputer for database input, batch data entry using in-house computer centers, and online data entry to Inforonics' mainframe computer. Two final sections deal with special procedures for encoding data backlogs when converting a large reference publication to a mechanized system, and the lack of satisfactory character and text coding standards. It is recommended that authors generally adopt their vendor's coding standard in the absence of a publishing industry-wide standard. (ESR)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED244594

# S T A T E - O F - T H E - A R T  R E P O R T

## DATA INPUT FOR PUBLISHERS

Lawrence F. Buckland
President
Inforonics, Inc.

8 March 1984

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Lawrence F. Buckland
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR011115

STATE-OF-THE-ART REPORT

DATA INPUT FOR PUBLISHERS: Techniques for authors, editors, typesetters, and data base producers.

## INTRODUCTION

Publishers are rapidly adopting new methods of preparing manuscript which allow authors and editors to "set their own type." These methods use word processors, low cost microcomputers, intelligent data entry terminals, on-line computers, or optical character recognition (OCR) equipment to capture original manuscript on machine media. The obvious advantage of the new methods is reduced cost, because keyboarding does not have to be repeated by the typesetter. In addition, administrative tasks involved in sending and receiving copy, galley proofs, and other materials are eliminated when authors and editors control their manuscript. This advantage is often overlooked because it is not a direct cost associated with purchased typesetting services.

In its 21 years of existence, Inforonics has seen a shift in the publisher-typesetter interface toward in-house preparation of data files. A significant amount of the input, keyboarding, and data editing is performed by our customers at their location. This transition has not been without its problems, but we have learned a number of lessons and have developed a set of successful keyboard entry and edit procedures which customers use to enter data. The determination of the best input method depends on the size and number of jobs a customer processes. The purpose of this report is to provide our present and potential customers with an overview of what is possible by briefly describing each procedure and providing an estimate of processing costs. This report is intended to serve as a guide in selecting the most suitable input method. A final section of the report deals with special procedures for encoding backlogs of data when converting a large reference publication to a mechanized system.

## PROCEDURES AND EQUIPMENT FOR DATA ENTRY OF TEXT

The following sections describe data entry procedures and data editing procedures available from Inforonics. They are arranged in order of increasing equipment cost.

### 1. Optical Character Recognition (OCR)

Inforonics rarely uses OCR copy because the decreased cost of personal computers has made this new method more cost effective than OCR. Inforonics currently uses OCR only for converting previously typeset copy, described later in the report. The OCR procedure is described here to make the scope of the report complete, though somewhat historical.

The data input system which requires the simplest and least expensive equipment is optical character recognition. Using this technique, data is typed in a standard format on sheets of smooth-surface paper. These sheets are fed into the input hopper of an OCR scanner which converts them to computer form. Data is output on computer storage media, directly into a computer, or to a computer via a telecommunication

link.

The cost of the typewriter suitable for keyboarding OCR sheets is approximately $1,000.00 and the average industry cost for scanning is approximately $0.30 per 1,000 characters for clean OCR sheets. (Approximately $0.80 for data for a 6 x 9 book page.)

Some scanners have an editing capability which allows an editor to correct OCR sheets by marking. The scanner interprets the marks, making the specified corrections. In this procedure, an operator monitors the scanning process by correcting data which the scanner can't read and inserting data where necessary. Costs for this monitored type of scanning are considerably higher and can range up to $0.80 per 1,000 characters.

Although OCR scanning offers the advantage of the lowest equipment cost, significant difficulties can occur. These stem from possible scanner inaccuracies in recognizing the keyboarded characters on the OCR sheets. There are numerous reasons for errors, including improper clarity of type impression, ribbon and paper used, cleanliness of the scanner, alignment of the typescript on the OCR sheet, etc. It suffices to say that to be successful, all components of the OCR process have to be "finely tuned." This can be done easily if all components are in one location. As mentioned previously, Inforonics' in-house OCR system has been replaced by a more reliable microcomputer input system. Our previous experience has shown that the quality control of OCR work done off the premises cannot be maintained adequately, and therefore we never accepted this type of customer input.

2. Microcomputers and Home Computers

Prices are rapidly decreasing for computers assembled from large scale microcircuitry. These computers are now priced from $500.00 to $5,000.00. One sizeable share of this market is the so called "home or personal computer" which can be used as a data input device. Low cost programs are available for home computers which can be used to create text files. Inforonics presently accepts diskettes from a number of personal computers which range in cost from $2,000.00 to $5,000.00 depending on the number of disk drives and whether or not a printer is used. Diskettes are delivered from a customer's computer to Inforonics, where they are read and transmitted to our computer. The specifications for an ASCII file recorded on a personal computer diskette can be obtained by writing Inforonics, Inc. If Inforonics does not have a disk reader which can read the customer's diskette then the data must be accepted by data transmission via phone lines. When data is transmitted it costs $.40 per 1,000 characters or $1.10 for a 6 x 9 book page (plus, if applicable, regular phone charges). This service is the best method of data input we have used to date. Its equipment cost is less than word processors and minicomputers, and it is a more reliable input method than OCR.

The one caution that must be mentioned is that the software support for home computers may not be adequate for the publisher. Although there is software available for these machines, it is a mass market product and may not solve all of the aspects of a publisher's input problem. Inforonics and other text processing service companies accept disks to

a local standard specification; however, they leave the software setup needed to produce these disks up to the customer and the home computer "store." If data recording capability problems arise, Inforonics will solve the problem by providing software services on a consulting basis. Because such software consulting services usually require costly programming, its cost can approach or exceed that of the home computer if one is not careful.

3. Inforonics' supplied personal computer

To help our clients through the personal computer selection process we have chosen a computer and developed software called TexTerm to accomplish the data input function. The computer is sold or rented to the client at cost (approximately $5,000.00). Inforonics charges $1,500.00 for setup, test of the application, and training of the client's personnel in the task of keyboarding. (A complete description of Inforonics' TexTerm can be obtained by writing Inforonics.)

4. Word Processors

The general availability of word processing in the office environment offers the publisher another option for keyboarding text for publications. Word processors are micro- or mini-computers with special purpose software for keyboarding, editing, and printing text. These machines are used to produce all types of office documents ranging from correspondence to manuals and reports. More recently word processor manufacturers are expanding their functions to simple accounting and file search so that they are becoming "office automation" machines. Costs range from approximately $7,000.00 for a single terminal system to $30,000.00 - $150,000.00 or more for a multi-terminal shared logic system.

Inforonics accepts data from word processors and enters it into its TPS processing system at a cost of $0.40 per 1,000 characters or $1.10 per 6 x 9 book page. The primary advantages of word processing are equipment availability and superior editing capability. However, problems with word processing occur because of incompatibility of the storage diskettes used in different brands of word processor. This diskette incompatibility could mean that a typesetting service vendor such as Inforonics must have several types of diskette drives and many different sets of disk reading software to accept and read all types of diskettes. At Inforonics this approach was thought too awkward and too expensive. We have chosen to standardize input by accepting word processing data via telecommunication. This requires that the word processor have transmission capability with one of the industry standard communication protocols and that the publisher use a standard printwheel character set.

Currently, Inforonics accepts TTY and BISYNC protocols. The TTY protocol which is most common on word processing machines, does not have error checking beyond parity. Thus, the telephone transmission must be reliable, or errors will creep into the data. The long range solution to reliable word processing transmission is to use error correcting methods; users, when selecting a word processor, should question the manufacturer on this point. It should be pointed out that very low-cost word processing programs are available to work on personal computers at a cost of computer plus program for less than the

dedicated word processors.

(For an account of one customer's use of a word processor for text input see Lawrence F. Buckland, "Book Composition by Authors--System and Equipment Considerations," in the 1979 SSP Proceedings, pp 49-53, Washington: Society for Scholarly Publishing, 1980. Reprint available from Inforonics, Inc.)

(For a description of keyboarding specifications to create word processing data to be transmitted to Inforonics, send for Inforonics' Technical Memorandum, "A New Way to Set Type from Word Processors.")

5. Microcomputer Data Base Input

A better alternative to a word processor, is a programmed microcomputer. Rather than adding additional software capability to a word processor to make it more versatile, a small, general purpose microcomputer is programmed to handle the data entry task. Several vendors can supply programmed data entry systems including Inforonics. This type of equipment and its associated software extend the data entry concept to achieve field prompting and data validation. Field prompting is the capability of the system to ask the typist for manuscript items by automatically supplying the identifiers, or field tags,assigned to each item. The terminal accomplishes this by displaying a "work sheet" tailored to the individual manuscript. This feature allows input from raw source data eliminating the need to edit or transcribe data onto clean manuscript or data creation sheets. Tne second function, data validation, is the capability of the system to inform the typist (with a beeping sound) that a typing error has just been made. Thus, the system catches a majority of the typing or editing errors before the manuscript leaves the operator's hands.

The data base input system also has a printing capability to produce copy for proofreading, an editing capability to correct or change copy, and a transmision capability with error detecting and correction functions. The edit function has the full data validation capability of the keyboard input mode so that errors are also detected while correcting copy.

Including software, the cost of this system ranges from $5,000.00 for a single terminal system to $40,000.00 for a multi-terminal system. It should be remembered that the micro-computer may also perform other data processing functions such as mailing list maintenance and accounting. The cost for data transmitted to Inforonics from this terminal is approximately $0.40 per 1,000 characters or $1.10 per 6 x 9 book page. (A complete description of Inforonics' TexTerm can be obtained by writing Inforonics.)

6. Data Entry Using In-House Computer Centers

Customers can use in-house data processing centers to prepare text and deliver data to Inforonics on magnetic tape. Data is entered into these systems using punch cards, key to diskette, key to disk, key to tape, or an on-line data entry and edit system. Once entered into the computer, data can be printed out, proofread, and edited. Complete correct files can be written out on magnetic tape and delivered to Inforonics. Two types of formats can be accepted by the Inforonics

system. The preferable format is industry ASCII 1600 BPI magnetic tape recorded in a variable length format with fields tagged according to the data structure of the manuscript, one physical record to one logical record. There are some conventions for representing character sets extending beyond standard ASCII which need to be followed if the manuscript contains special symbols. A complete description of the ASCII magnetic tape format can be obtained by writing Inforonics, Inc.

The second tape format accepted by Inforonics is industry compatible EBCDIC in which each item in the manuscript is assigned a fixed field position in the record. A description of the EBCDIC format can be obtained by writing Inforonics.

Magnetic tape from a customer computer center is a very reliable low cost method for provision of information to Inforonics as long as the conventions are followed. The cost for this data entry is $0.06 per 1,000 characters or approximately $0.15 per 6 x 9 book page.

7. On-Line Data Entry

The last method of data input for publishers used by Inforonics is text entry on-line to Inforonics' computer. The method allows rapid data entry, immediate printout at the user's terminal, and on-line editing to remove typographic errors or to change text. Although the system is flexible, it is costly because of the telecommunication costs and computer resources used. The charges are made on a connect hour basis at approximately $20.00 per hour. This method on the average costs $2.00 per 1,000 characters.

This on-line input method is rarely used for publication files except for small last minute additions or corrections. In the latter application the costs are low because a small amount of data is entered. The costs are easily justified because the rapid communication capability of the method greatly speeds up the production process in the final stages of composition.

## SPECIAL INPUT PROBLEMS - CONVERTING BACK FILES

Users have a special problem when converting manually prepared publications to data base publications. The problem of conversion of the back file to machine form has several facets. First, it requires a large amount of keyboard labor compared to that required for the ongoing publication. Second, the data is rarely in the form needed for input. For example, field tags may not be present, or fields added in the data base design may be missing or appear in a separate section of the manual publication such as an index.

Inforonics offers several solutions to this problem: (1) keyboarding may be handled by Inforonics; (2) keyboarding may be performed overseas; or (3) a more powerful OCR scanner may be used to read previously typeset copy. In keyboarding, the field tagging is done while typing. If scanning is used, field tagging is done by a complex recognition program which processes the scanned output.

### 1. Keyboarding by Inforonics

For large back files of source data or pages from previous editions of a publication, Inforonics will keyboard the data. Our typists will interpret the data, determine the field tags, and insert them while typing. The data will be printed out, proofread, and corrected. For normal copy, the cost for this input ranges from $3.15 per 1,000 characters to $6.00 per 1,000 characters depending on the complexity of the copy and on whether Inforonics or the customer does the proofreading.

### 2. Overseas Keyboarding

An equivalent service to the above can be provided using overseas keyboarding labor. The cost of this service, including Inforonics' supervision, is $1.50 per 1,000 characters for keyboarding, proofreading, and correction.

### 3. OCR Scanning

There are several service bureaus which employ Kurzweil scanners. (Kurzweil is a subsidiary of Xerox Corporation.) This scanner can read typeset pages with considerable accuracy. The resulting file is computer processed to edit it to a field tagged form. This system is well-suited to directories and other simple material. Complex text requires a human edit pass to detect errors and supply field tags for ambiguous typographic forms read by the scanner. Inforonics will subcontract this service and charge the customers between $0.80 and $1.50 per 1,000 characters depending on the complexity and size of the conversion tasks.

## CHARACTER AND TEXT CODING STANDARDS

Although the main thrust of this report is to describe and give advice about input procedures and equipment, we would be remiss in our position if we did not say something about data encoding standards. There are 6 aspects to the standard situation today.

1. The current computer and telecommunication standards: ASCII (American Standard Code for Information Interchange) and EBCDIC (Extended Binary Coded Decimal Interchange Code) are not satisfactory for publishing because they do not allow one to represent the characters needed or the typographic layout publishers need.

2. There are no standards of code adequate for the publishing industry.

3. Data base standards such as are used for indexes and directories are different from data encoding standards for typographic formatting.

4. AAP is currently contracting with ASPEN Systems to do the necessary survey and analytical work to arrive at a proposed standard, but the adoption of such a standard is at least two years away.

5. Various vendors have developed their own standards in order to make their own processing more efficient.

6. Vendors will adapt to accept your standard if your work volume is high enough.

Considering the situation described above, the best advice is to adopt the standard convention of your data base processing or typesetting vendor, unless you have a large volume of work in which case you may achieve economies by developing your own. The current problem with converting "home grown" floppy disks is the cost of adapting the vendor's system, not a fundamental technical incompatibility. If you do not have experience in coding systems in data base design, engage one of the vendors to help you. They have done hundreds of jobs and are bound to have experience far more extensive than yours. Once developed the standard and procedure can be installed on the particular data input device you have chosen.

SUMMARY

Beginning with newspaper publishers, more and more publishers are having authors and editors prepare their manuscripts in machine form, so that they can be converted directly to type. The number of methods for creating this machine form manuscript are numerous, and each has its capabilities and advantages.

The OCR method, while its keyboard equipment is cheapest to acquire, is too error prone to be useful in every-day production. Where keyboards are remote from the scanners the OCR image impression quality is difficult to control.

Word processors, if already available to an author or editor, are most cost effective for low-volume use. No additional equipment cost is required, and they offer reliable, easy input and editing capability.

Pre-programmed microcomputers sold by Inforonics and other systems vendors are best because they can be tailored to make a publisher's particular input keyboarding easy and fast, justifying their cost for high-volume use. Clever users or users with access to programmers can save equipment costs by programming their own micro- or home computers to accomplish a data input task. Finally, if large amounts of backlog data are to be converted, overseas keyboard services offer fast, low-cost capability.

Finally, unless you have a large volume of work, and until a publishing industry-wide standard is developed, adopt the coding standard of your vendor. The vendor will help you and it is easier for him than trying to convert a disk typed according to a unique procedure developed by yourself.

LFB:mek